How to Commit the Gambler's Fallacy and Get Away With It

Davis Baird

University of Arizona

and

Richard E. Otte

University of Arizona and University of Pittsburgh

1. The Gambler's Fallacy

The gambler's fallacy runs like this: We know

1) the bias for heads of a certain coin and flipping device is 1/2, and

2) flips of the coin with the flipping device are independent.

Given these facts we can easily compute the probability of 4 heads or fewer turning up in 20 flips, $P(h \leq 4)$, to be

$$(20!(1/2)^{20})/\Sigma_{i=0}^{4}(i!(20-i)!) \cong .0059.$$

Suppose we flip the coin 15 times and get--improbably enough--all tails. If the coin lands tails up once more in the next 5 flips the improbable event of 4 or fewer heads in 20 flips, $h \leq 4$, will have occurred. Since $P(h \leq 4)$ is so small we might be tempted to infer that all of the next five flips will be heads. If we did not wish to make the strong claim that all of the next five flips will be heads, we might at least be tempted to claim that the probability of heads on each of the next five flips is considerably greater than 1/2. Thus the gambler betting on red at a roulette table might exclaim after a string of bad luck--15 black numbers in a row--"If this wheel is fair the next number is bound to be red!"

Such reasoning is fallacious. To determine $P(h \leq 4)$ we assume that flips of the coin are independent of each other: the outcome of past flips does not alter chances on future flips. The probability of

PSA 1982, Volume 1, pp. 169-180

heads on the 17th flip is simply 1/2, regardless of whether the first 16 flips were all heads, all tails, or any combination of heads and tails. Indeed, the probability of getting 5 heads in a row on the 16th, 17th, 18th, 19th and 20th flips is quite small:

$$(1/2)^5 \cong .03.$$

It is not large, as those reasoning a la gambler's fallacy conclude.

This much is uncontroversial, and we shall not dispute it. There are, however, patterns of reasoning very much like the gambler's fallacy which we believe are not fallacious. We provide examples of such patterns of reasoning and argue that they demonstrate that probabilistic/inductive reasoning is very sensitive to the statistical model. In particular we exhibit two different statistical models for coin flipping which are mutually consistent, yet which license contradicting inferences; on one model the gambler's fallacy is a fallacy, and on the other it may be a legitimate mode of inference.

2. Emergent Probabilities

Our point of departure is a recent paper by Ian Hacking, "Grounding Probabilities from Below" (Hacking 1980). Hacking begins by asking:

> Is it the case that every stable frequency, correctly represented by a mathematical probability, is 'grounded from below' by probabilities that apply to individuals? That is, does the frequency distribution in the population derive from probabilistic facts about the individuals that compose it? Or are there some stable frequencies that pertain to populations, but do not derive from probabilistic facts about members of the population? (Hacking 1980, p. 110).

Hacking presents a case for the 'emergentist position': there are probabilistic laws, manifested by stable frequencies in populations, where no probabilistic laws cover individuals in the population. Put another way, Hacking thinks that there may be cases where probabilities can be correctly assigned to ensembles while no probabilities can be correctly assigned to individual members of the ensembles in question. Hacking calls such probabilities, probabilities which are not 'grounded from below'.

Much of Hacking's argument rests on two examples: suicide statistics from 19th Century Europe, and fertility statistics from Germany during the period 1871-1939. The interesting feature of both of these cases is that stable long run frequencies are found in populations while no stable frequencies are found in partitions of these populations. Concerning fertility statistics, Hacking writes:

> Knodel found that the national decline in birth rate is nicely mirrored in each administrative district or Kreis. The numbers are as regular as any which could be hoped for in demography. But,

> when we pass to smaller units, such as the village, the uniformity collapses. Although every Kreis is doing the same thing as every other, villages within Kreise are all going their own ways, without much in the way of underlying laws. (Hacking 1980, p. 114).

Hacking concludes from this collapse of uniformity that:

> [T]he propensity to limit family size may simply not be a number which represents a property of each of the individual couples within the Kreis. There is only a very striking property which applies to the Kreis as a whole, that the birth rate is declining in a systematic and law-like way. (Hacking 1980, p. 115).

When a statistical uniformity appears in a population, there are two ways to account for it. One way is to ascribe probabilistic properties to individuals and use results such as the law of large numbers to explain stable regularities in ensembles of individuals. Another way is to claim that the uniformity in the population does not arise out of any probabilistic facts about the individual members of the population, but rather that the probabilities are manifest only at the level of an ensemble.

We find Hacking's paper suggestive, and shall tentatively grant him his major point. There are cases where probabilistic properties may be correctly ascribed to individuals, and, by inferences based on the probability calculus, to populations; and there are cases where probabilistic properties may be correctly ascribed only to ensembles in which no probabilistic properties may be ascribed to individuals within the ensemble. However, we take issue with Hacking on one of his closing remarks: "I do not think my distinction makes the slightest difference to any practical problem of statistical inference." (Hacking 1980, p. 115). While it is not entirely clear which problems are practical and which are not, we believe that Hacking's two cases license different inferences.

3. A Trivial Example

Consider the following fiction. We are gathering census data in Massachusetts. We happen to be aware of a well established statistical law that the number of children born per couple--on average--in any county is 2.44. Let us further suppose that this statistical law is not grounded from below: no probabilistic properties ascribable to individual couples or even individual towns grounds the statistical law. We have completed the census for every town in Middlesex County except Lexington. The following facts have been revealed:

1) There are 980,000 couples in all of Middlesex County except Lexington;

2) There are 2,401,000 children in all of Middlesex County except Lexington.

We also estimate that:

3) There are 20,000 couples in Lexington.

From 1) and 2) it is easy to compute the average number of children per couple in all of Middlesex County except Lexington:

2,401,000/980,000 = 2.45.

Since it is a statistical law that counties have stable average family sizes, while towns may have widely varying average family sizes, we predict the average number of children per couple in Lexington from that data. Let X be the number of children in Lexington. Then our statistical law requires that:

(X + 2,401,000 / 980,000 + 20,000) = 2.44.

Hence,

X = 2.44(980,000 + 20,000) - 2,401,000 = 39,000.

Accordingly we predict that the average number of children per couple in Lexington is

39,000/20,000 = 1.95.

This is different from the county average of 2.44.

If the statistical law used above was grounded from below then the reasoning above would commit the gambler's fallacy. Assuming they exist, let us call the grounds for the statistical law, 'sub-laws'; these sub-laws might stipulate probabilistic facts about the average number of children per couple in the individual towns. All these sub-laws together allow us to compute that the average number of children per couple in Middlesex County should be about 2.44. The important point is that the County average would be derivative upon individual town averages. Thus, to draw the analogy with the gambler's fallacy on the roulette wheel: There is a long run statistical stability to the effect that only very infrequently will fewer than 5 red numbers turn up in 20 plays of the wheel. This is analogous to the county law, which also stipulates a long run statistical stability. The long run statistical stability of the roulette wheel is derivative upon the individual probabilities assigned to the individual numbers on the wheel and the independence of plays; in the same manner, the county law is derivative upon the 'sub-laws' concerned with the individual towns. When the gambler commits his fallacy he reasons about the remaining 5 plays on the basis of the long run statistical stability in the same way that we have reasoned about Lexington on the basis of the county law. The gambler commits a fallacy because his reasoning violates one of the assumptions used to compute the law concerning the long run statistical stability. Similarly, we violated one of the sub-laws grounding our county law in making our inference about

Lexington. If our county law is grounded from below, our reasoning commits the gambler's fallacy.

On the other hand, if the county law is not grounded from below by sub-laws about the individual towns, then our reasoning to a conclusion about Lexington does not violate any assumptions used in determining the county law. Consequently, in such cases it is not a fallacy to reason as we did above. Indeed, such reasoning is similar to a very common form of reasoning. We have a law that tells us that planets orbit in ellipses. We gather some data concerning the positions of a particular planet at some particular times, and we fit an ellipse to the observations--as our law bids us. The ellipse then allows us to compute where the planet will be at other times. We claim that when probabilities are not grounded from below, our reasoning concerning Lexington is analogous to the reasoning in the planet-ellipse example. It is not a fallacy at all.

4. A Coin Flipping Model for Emergent Probabilities

Under normal circumstances we assume the following about coin flipping models of probabilistic phenomena:

1) the coin (and perhaps the flipping device) has a property, called bias; the bias, Θ, of a coin [and flipping device] is the probability that the coin will land heads on any flip;

2) distinct flips of the coin are statistically independent of each other.

Let us call a model which has these properties M_b. Alternatively, we can define the following model about coin flipping type phenomena:

1') The coin (and perhaps the flipping device) has a property, called smias; the smias, σ, of a coin [and flipping device] is the probability that the coin will land heads five or more times in 20 flips;

2') distinct 20 flip trials of the coin are statistically independent of one another.

We will call such a model M_s.

If there is a M_b model then there is associated with it an M_s model. If a coin has bias Θ then the coin will have smias

$$\sigma = 1 - 20!\ \Sigma_{i=0}^{4}((\Theta^{i}(1-\Theta)^{20-i})/(i!(20-i)!).$$

Since individual flips are independent given a M_b model, 20 flip trials are also independent. Thus, given a M_b model we can find the associated M_s model.

If we admit the possibility of emergent probabilities, the converse need not be true. There may be cases where we can make probabilistic statements about the behavior of a coin on trials of at least 20 flips, while no probabilistic statements can be made about trials of fewer than 20 flips. Perhaps there is some unknown feature of the flipping device which keeps individual flips from being independent, but assures that each 20 flip trial is independent of other 20 flip trials.

Given our preconditioning to think of coins in terms of bias, our smias model appears a bit strange. However, such a model does provide a simple picture of what is going on with Hacking's emergent probabilities. It is perhaps worthwhile to dispel some possible objections by requiring of our smias model some criteria of identity for 20 flip 'individual' trials. By analogy then our individual 20 flip trials operate like Hacking's Kreise. We cannot make probability statements for partitions smaller than the Kreis, or a 20 flip trial, while we can do so for Kreise. Such is a situation where we have a smias model and no bias model to ground it from below.

Hacking says that such emergent probabilities make no difference to statistical inference, but they do. If there is an adequate bias model then reasoning about the remaining 5 flips of a twenty flip sequence on the basis of the previous 15 flips commits the gambler's fallacy. However, if there is no bias model 'underneath' a smias model, then such reasoning is sound. Given 15 tails in the first 15 flips of a 20 flip trial, if we assume smias of .999 and no bias underneath, we can legitimately infer that the next 5 flips will be heads. This is how to commit the gambler's fallacy and get away with it.

5. Bias, Smias and the Theory of Statistical Tests

Our smias model of coin flipping might appear peculiar. On the contrary, however, we believe that smias or something similar is in fact the model tested when statistical tests are run on hypotheses about bias. To clarify this paradoxical remark: we frame a hypothesis in terms of bias; the test of this hypothesis, however, concerns only properties of the smias model associated with the bias model purportedly under test. While the two possible results of a statistical test of some hypothesis about bias is said to be 'reject' or 'do not reject', all we are really justified in inferring is the rejection or not of the associated hypothesis about smias. An example makes these remarks clearer.

Suppose we wish to test the hypothesis H that a coin has bias greater than 1/2,

$$H: \ \Theta \geq 1/2.$$

According to the standard theories of statistical testing of either Neyman/Pearson or Fisher, to test we adopt a rejection region, R. The rejection region is a set of possible outcomes; if any of these possible outcomes occurs we reject the hypothesis. Suppose in the case at hand we can afford to flip the coin only 20 times. Obviously the rejection region, R, should include those results where a preponderance of tails came up. The rejection region, R= {0,1,2,3 or 4 heads in 20 flips}, provides a fairly stringent test of H. Stringency, or level of significance, is defined as the probability, assuming the hypothesis under test, of the rejection region; in this case the probability of R is about .0059.[1] In a certain sense stringency is the probability that such a test would lead to a mistaken rejection of a true hypothesis. In this case that probability is small, 59/10,000. It is important to note that in order to perform such a test we need only know whether or not a result occurred in the rejection region. According to accepted testing theory it is the information of whether or not R is true, and not the actual result, that is important for inference.

Alternatively, suppose we wish to test a hypothesis J about smias:

J: $\sigma \geq .9941$.

Our financial constraints remain the same; we can afford only one 20 flip trial. Since we can perform only one 20 flip trial, only two outcomes are possible: 'success', 5 or more heads in 20 flips, or 'failure', 4 or fewer heads in 20 flips. Obviously the sensible test rejects J if and only if the result is 'failure'. Such a test is quite reasonable; like our test for bias above, it has stringency of .0059.

Obviously the test of H is identical with the test of J. We reject either hypothesis, H or J, just in case 4 or fewer heads turn up in 20 flips. In a certain sense this is just what we want. If there is a bias model, then

$H = \Theta \geq 1/2$ if and only if $\sigma \geq .9941 = J$.

However, if there is no bias model in the offing, then the equivalence fails. But we should note that even if there is no bias model we can still, in a purely formal sense, test hypotheses about a non-referring parameter, bias. The only features of the bias model which are tested are features of the smias model associated with it. When we test hypotheses about bias we are really concerned only with smias properties.

One might object: The only reason our test of bias is identical to a test of smias is that we could afford to flip the coin, conveniently enough, only 20 times. This is true, however the important point is that no matter how many times we do flip the coin, we can always define a smias type model for the appropriate number of flips. Our test of bias will always be equivalent to a test of an appropriately defined hypothesis in the associated smias type model.

Our smias/bias example plays on a well known problem: why is the stringency of a test measured by the probability of the entire rejection region? Should it not be measured by the probability of the specific result, e.g., 3 heads in 20 flips, instead of simply R? Recall that in order to test H, $\Theta \geq 1/2$, we had only to know whether or not R occurred. While several proposed justifications for this procedure have appeared, none are entirely satisfactory. We believe considerations about smias, bias, and emergent probabilities shed some light on this problem.

There is reason to believe test stringency should not be measured by the probability of the rejection region. An almost universally adopted maxim of inductive inference is that the premises of an inductive argument must include <u>all</u> of the <u>relevant</u> known information concerning the conclusion of the argument. This requirement of total evidence was perhaps first made clear by Rudolf Carnap in his 1947 paper on inductive logic; a recent discussion with particular attention to problems of statistical inference is in Seidenfeld 1979. The most widely accepted view is that what are called 'minimally sufficient statistics' contain all the relevant information for inference (Seidenfeld 1979, p. 83). Unfortunately, whether or not the outcome occurred in the rejection region is not sufficient information for inferences about bias. Hence, according to the sufficiency reading of the requirement for total evidence, the occurrence or non-occurrence of the rejection region is not sufficient information for inferences about bias. It seems then, that measuring stringency by the probability of the rejection region is not right.

While the occurrence of R is not sufficient for inferences about bias, the occurrence of 'failure' is sufficient for inferences about smias. R is identical to 'failure'. Hence, it seems that the information collected in a test of bias is really only sufficient for inferences about the associated smias hypothesis. On the standard theory of testing, we test bias by testing the associated smias.

If bias is a property of the coin, then we might justify the rejection of hypotheses about bias, such as H, on the basis of insufficient information as follows: The information that a result in the rejection region occurred is sufficient for hypotheses about smias. In particular, if R occurs we have good reason to reject J, $\sigma \geq .9941$. Since we assume our smias probabilities are grounded from below with bias probabilities, we infer that H if and only if J. Hence, if we reject J, we should also reject H.

6. Metaphysics vs. Epistemology

We believe that the tension between choosing a bias model and a smias model reflects a tension that exists between metaphysical intuitions and epistemological intuitions. Russell states the metaphysical intuitions with characteristic flair:

> The theory of probability is in a very unsatisfactory state, both logically and mathematically; and I do not believe that there

> is any alchemy by which it can produce regularity in large numbers out of pure caprice in each single case. If the penny really chose by caprice whether to fall heads or tails, have we any reason to say that it would choose one about as often as the other? Might not caprice lead just as well always to the same choice?... [W]e cannot accept the view that the ultimate regularities in the world have to do with large numbers of cases, and we shall have to suppose that the statistical laws of atomic behavior are derivative from hitherto undiscovered laws of individual behavior. (Russell 1935, p. 168).

Russell explicitly states here that the emergentist position is unsatisfactory. Russell believes, for metaphysical reasons, that we must explain statistical regularities by an appeal to deterministic laws about individuals; statistical laws are dependent upon such laws. For example, consider the statistical laws of classical kinetic theory. These laws are based upon deterministic laws about individual molecules. Their statistical character arises only because of the difficulty in stating the initial conditions of a system and working out the appropriate calculations. Evidently Russell believes that all statistical laws should be explained in this manner.

Russell wants a 'deeper' deterministic explanation for long run stability. There are two parts to his worry. One is a general uneasy feeling regarding indeterministic hypotheses. The other is a belief that laws operate at the level of individuals--not collectives. More current sentiment regarding determinism is well expressed by Salmon:

> This view involves an _a priori_ commitment to determinism--... . This position seems to me untenable. I do not mean to argue that present physical theory is complete and correct, but, rather that there is no reason to make an _a priori_ decision as to the nature of further physical theories. Perhaps, in the future, improved theories will provide a deterministic account of events that current theory regards as causally undetermined--but perhaps they will not. We should be prepared for the possibility that the indeterministic character of physical theory is correct and that there are events which are intrinsically improbable, not merely improbable in relation to our present incomplete knowledge. (Salmon 1971, p. 9-10).

We agree with Salmon, and think his attitude towards determinism is also appropriate for the second part of Russell's worry. It is an empirical question as to what level natural laws operate at.

In our opinion the level at which laws operate should be left open and not assumed _a priori_ to be that of the individual. This is not generally accepted, as Hacking's reference to the 'much maligned topic of emergentism' indicates (Hacking 1980, p. 110). While we seem to have outgrown our deterministic prejudices, as the acceptance of Quantum Mechanics and the growing popularity of probabilistic theories of causality shows, we remain clinging to the second half of

Russell's worry: laws operate at the level of individuals. The result is a metaphysical 'push' away from properties like smias towards bias and beyond.

On the other hand, there is an epistemological reason to be unhappy with a law such as $\theta = 1/2$. Such a law is completely untestable, except in terms of its consequences for many flips. Suppose that we flip a coin once and it comes up heads; does that give us any reason to either claim the coin is fair, or unfair? Of course not. We accept or reject the hypothesis on the basis of many flips. But in dealing with many flips, we are dealing with smias-type properties. It seems like this gives an epistemological reason to work with smias-type models, and not be concerned with properties such as bias. In a sense then, there is an epistemological 'push' towards smias away from bias.

There appear to be three levels of hypotheses that we are dealing with in this discussion: deterministic laws, bias type properties, and smias type properties. We might sketch the situation as follows:

↑	Deterministic Laws	
	explain	
metaphysical push	Bias type properties: $\theta = 1/2$	epistemological push
	explain	
	Smias type properties: $\sigma = .9941$	↓

The result is confusion. We have a problematic theory of testing where we do not use all the relevant information when inferring about bias; and we have an inadequate explanation of the regularities observed in ensembles by means of untestable hypotheses about bias.

7. Is the Gambler's Fallacy Really a Fallacy?

We simply are in a position of not knowing for certain whether smias properties are always, ever, or never grounded from below by bias properties. In some cases we may be able to gather some pertinent data. We can test for independence of flips. Such tests are not, however, conclusive. Ultimately our belief in grounding probabilities from below rests upon a metaphysical assumption. This intuition may be very plausible but it is metaphysical, nonetheless. Consequently, we can never be certain whether the gambler's fallacy is really a fallacy, since it also rests upon this metaphysical assumption. Perhaps there is some consolation in this. After all, how could so many gamblers be so wrong?

The later part of our exposition has been given in terms of bias and smias. There are two reasons for this. In the first place, it is helpful to have a concrete model in mind; coin flipping serves nicely in this role. It should be clear though, that smias type probabilities throughout stand for any sort of population probabilities, and bias type probabilities stand for the possible grounds from below that the smias type probabilities might enjoy. Secondly, Hacking discusses binomial coin-like phenomena as cases where Poisson's analysis shows how emergent regularities can be derived from the probabilistic properties of the individuals that compose the population. (Hacking 1980, pp. 112-3). Hacking closes by asking for Poisson's analysis to be extended:

> A useful continuation of Poisson's programme for a law of large numbers would examine the extent to which quantitative frequencies within a collective may derive from purely qualitative and unstructured propensities assigned to individual members of the collective. (Hacking 1980, p. 115).

We second Hacking's proposal for a continuation of Poisson's programme. We would add, however, a request for a better explanation of why we ought to adopt bias-type models even in those cases considered by Poisson.

Notes

[1]This exposition is condensed. The theory of testing is not our principal concern here and we have accordingly not discussed it in detail. Those interested may consult standard texts of Neyman (1952), or Fisher (1956 and 1970), and Cox and Hinkley (1974). Philosophical work on hypothesis tests may be found in Hacking (1965), Seidenfeld (1979), and Baird (1981).

References

Baird, D. (1981). Significance Tests: Their Logic and Early History. Unpublished Ph.D. Dissertation, Stanford University.

Carnap, R. (1947). "On the Application of Inductive Logic." Philosophy and Phenomenological Research 8: 133-148.

Cox, D.R. and Hinkley, D.V. (1974). Theoretical Statistics. London: Chapman and Hall Ltd.

Fisher, R.A. (1956). Statistical Methods and Scientific Inference. New York: Hafner Publishing Company.

-----------. (1970). Statistical Methods for Research Workers. 14th ed. New York: Hafner Publishing Company.

Hacking, Ian. (1965). Logic of Statistical Inference. Cambridge: Cambridge University Press.

------------. (1980). "Grounding Probabilities from Below." In PSA 1980, Volume 1. Edited by P.D. Asquith and R.N. Giere. East Lansing, Michigan: Philosophy of Science Association. Pages 110-116.

Neyman, J. (1952). First Course in Probability and Statistics. New York: H. Holt and Company.

Russell, B. (1935). Religion and Science. London: T. Butterworth.

Salmon, W. (1971). Statistical Explanation and Statistical Relevance. Pittsburgh: University of Pittsburgh Press.

Seidenfeld, T. (1979). Philosophical Problems of Statistical Inference: Learning from R.A. Fisher. Dordrecht: Reidel.